The SDSS-GriPhyN Challenge Problems: Cluster Finding, Correlation Functions and Weak Lensing

James Annis[1], Steve Kent[1], Alex Szalay[2]

[1]Experimental Astrophysics Group, Fermilab
[2]Department of Physics and Astronomy, The Johns Hopkins University

Draft v1 November 7, 2001

We examine the Sloan Digital Sky Survey data pipelines in the light of virtual data. After a brief review of the SDSS project, pipelines, and data products, we consider the intersection of GriPhyN interests and SDSS activities. We then suggest that the appropriate prototype virtual data set is a weak lensing map derived from the co-added Southern Survey of the SDSS (where coadd is the process of adding images together). The request for the weak lensing map is a request for a catalog made from the coadded images, the existence or non-existence of which determines the need for running all of the SDSS pipelines, running special coadding code, and re-running the SDSS pipelines. While astronomers naturally focus on the image processing step of the coadd, the bulk of the process is dealing with the same multi-faceted bookkeeping problem that faces all Virtual Data applications. The creation of the coadded South is a challenge that the SDSS collaboration will be solving over the next two years.

VDT folks want over the next 6 months designs for testbeds. Designs and initial implementations.

The idea of the GriPhyN Google has power. As does Moore style metadata.

And aim this as the design document for the pair of Tier-2 centers.

Is the concept of a scan machine for the imaging data of relevance?

1 Introduction

The Sloan Digital Sky Survey is a project to map one quarter of the night sky. We are using a 150 Megapixel camera to obtain 5-bandpass images of the sky. We then target the 1 million brightest galaxies for spectroscopy, allowing us to produce 3-dimensional maps of the galaxies in the universe out to 5 billion light years. The final imaging mosaic will be 1 million by 1 million pixels.

Astronomers based at the Apache Point Observatory near White Sands, New Mexico use a specially designed wide field 2.5m telescope to perform both the imaging and the spectropscopy, and a nearby 0.5m telescope to peform the imaging calibration.

1.1 The SDSS Data Sets

The SDSS, during an imaging night, produces data at a 8 Mbytes/s rate. Imaging nights occur perhaps 20-30 nights per year and spectroscopy occupies most of the rest of the nights not dominated by the full moon. Nonetheless, imaging data dominates the data sizes.

The fundamental SDSS data products are shown in table 1. The sizes are for the Northern Survey and the Southern Survey will roughly double the total amount of data.

Table 1: The Data of the SDSS

Data	Description	Data Size (Gigabytes)
Catalogs	Measured parameters of all objects	500
Atlas images	Cutouts about all detected objects	700
Binned sky	Sky after removal of detected objects	350
Masks	Regions of the sky not analyzed	350
Calibration	Calibration information	150
Frames	Complete corrected images	10,000

Our reduction produces complicated data. The catalog entry describing a galaxy has 120 members, including a radial profile. If a different measurement is needed, one can take the atlas images of the object and make a new measurement. If one desires to look for objects undetected by the normal processing, say low surface brightness galaxies, one can examine the binned sky. And one is always free to go back to the reduced image and try a different method of reduction.

The SDSS data sets are representative of the types of data astronomy produces and in particular the types that the NVO will face.

1.2 The Challenge Problems

The SDSS data allow a very wide array of analyses to be performed. Most involve the extraction of small data sets from the total SDSS data set. Some require the whole data, and some of these require computations beyond what is available from a SQL database. We have chosen three analyses to be SDSS challenge problems: these highlight the interesting domain problems of catalog versus pixel analyses, of high computation load, high storage load, and balanced analyses.

2 Astronomical Virtual Data

2.1 The Data Complexity

TsObj

Atlas images, binned sky

Spectroscopy

SX versus flat files

2.2 Metadata: Design for Speed

The metadata requirements for SDSS catalog very naturally map from the concepts of the FastNPoint codes of Andrew Moore and collaborators. In this world view, Grid Containers are not files or objects, but nodes of a kd-tree (or perhaps some other tree structure with better data insertion properties). In this view what matters for performance is the ability to know what is in the container without having to actually read the contents. Consider a metadata example listing the most useful quantities in astronomy:

Ra, Dec	position on sky	bounding box
Z	redshift	bounding box
r	r-band brightness	bounding box
g-r	g-r color	bounding box
r-i	r-i color	bounding box

Given just those quantities in the metadata catalog, the execution time for a range search can be brought down from N^2 to N log N. The central ideas are exclusion (if the range to be searched for does not cross the bounding box, one need not read that container) and subsumption (if the range to be searched for completely contains the bounding box, one needs the entire catalog, again not reading the container).

Furthermore, there are great possibilities for speed up if one is willing to accept an approximation or a given level of error. Clearly the majority of time in the range search above is spent going through the containers that have bounding boxes crossing the search; it is also true that often this affects the answer but little, as the statistics are dominated by the totally subsumed containers. Having the relevant metadata in principle allows the user to accept a level of error in return for speed.

Often what one is doing is to compare every object against every other object. The tree structure above gives considerable speed up; another comparable speedup is allowed if the objects of interest are themselves in containers with metadata allowing the exclusion and subsumption principles to operate.

These considerations also suggest that a useful derived dataset will be tree structures built against the Grid containers, with the relevant metadata built via time-consuming processing but then available quickly to later users.

2.3 Derived Data: Design for Variations

Most of astronomy is derived datasets. One of the clearest examples is the identification of clusters of galaxies. Nature has made a clean break at the scale of galaxies: galaxies and entities smaller are cleanly identifiable as single entities; above galaxies the entities are statistical. Given that, there are many different ways to identify clusters of galaxies. The SDSS currently is exploring 6 different cluster catalogs.

Table 2: SDSS Cluster Finders

Cluster Catalog	Description	Data
MaxBcg	Red luminous galaxies	Imaging catalog
Adaptive Matched Filter	Spatial/luminosity profiles	Imaging catalog
Voronoi	Voronoi tessellation	Imaging catalog
Cut and Enhance	Spatial/color search	Imaging catalog
C4	Color-color space	Imaging catalog
FOG	Velocity space overdensities	Spectroscopic catalog

Each of these catalogs are derived data sets. They may, in principle, be downloaded for existing regions, or the algorithm may be run at individual points in space, or a production run of the algorithm may be scheduled. It is worth pointing out that

a. Each algorithms have changeable parameters,
b. Each algorithm evolves and hence has version numbers,

c. The underlying data can change as the reduction or calibration is re-performed.
We thus point out that versioning and the associated bookkeeping is important.

Finally, we note that generically in astronomy one wishes to attach derived data to the underlying data. Here cluster finding is not a good example, and we will turn to the environment of individual galaxies. View the underlying galaxy data as a table; what astronomers generically wish to do is to add columns. Examples include counting red galaxies in some raidus about each galaxy, counting all galaxies in some radius about each galaxy, summing the H-alpha emission from all galaxies with spectra in some radius about each galaxy, etc. The reigning technology in the SDSS is tables with row to row matching by position.

3 The SDSS Challenge Problems

3.1 The SDSS Challenge Problem 1: Cluster Catalog Generation

The finding of clusters of galaxies in the SDSS galaxy catalog is both a good example of derived data and a good place to try to start integrating GriPhyN tools with astronomical codes.

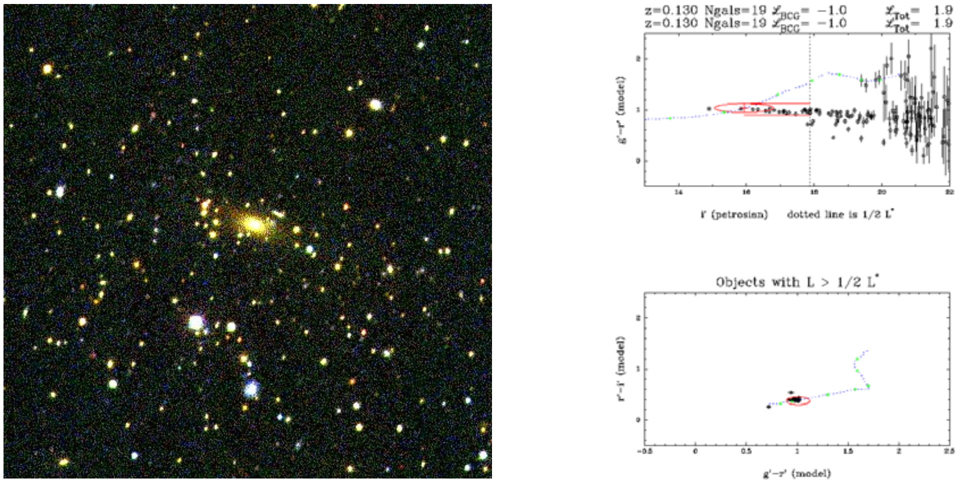

Figure 1: A cluster of galaxies seen in a true color image on the left, and as a color-magnitude and color-color plot on the right. The plots on the right illustrate one cluster finding technique, a matched filter on the E/S0 ridgeline in color-luminosity space: the number of galaxies inside the red is the signal.

Clusters of galaxies are the largest bound structures in the universe. By counting them at a variety of redshifts as a function of their masses, one is able to probe the evolution of structure in the universe. The number of the most massive clusters is a sensitive measure of the mass density Ω_m ;

combined with the Cosmic Microwave background measurements of the shape of the universe, these become a probe of the dark energy (aka, Cosmological Constant).

The basic procedure to find clusters is to count the number of galaxies within some range about a given galaxy. This is an N^2 process, though with us the use of metadata stored on trees it can be brought down to a N log(N) problem. Note that the procedure is done for each galaxy in the catalog.

The problem is computationally expensive, though balanced with the I/O requirements; with the appropriate choices of parameters it can be made either an I/O bound problem or a CPU bound problem. The problem faces moderate storage problems: a hundred square degrees of SDSS data masses to 25 Gig. The problem can be made embarrassingly parallel as there is an outer bound to the apparent size of clusters of interest. The work proceeds through many stages and through many intermediate files that can be used as a form of checkpoint.

The problem is a good choice for the initial challenge problem as 1) cluster catalogs are a very good example of derived data, 2) cluster catalog creation is roughly compute and storage balanced, 3) it can be solved in interesting times on existing testbeds, 4) it exercises the GriPhyn toolkit for solving problems of derived data catalog, replica catalog, and transformation catalog including DAG creation. As a side challenge, the existing cluster finding code requires exercising GriPhyn toolkit code migration tools.

3.2 The SDSS Challenge Problem 2: Spatial Correlation Functions and Power Spectra

The correlation function of the positions of galaxies projected on the sky, termed the angular correlation function, forms a Fourier transform pair with the spatial power spectrum. The power spectrum itself is of great interest in so much as the light from stars in galaxies traces the underlying mass both of normal matter and of the dark matter. If light traces mass, then when one measures the power spectra of galaxies one is measuring the power spectra of mass in the universe, a quantity that can be predicted quite accurately from the CMB fluctuations, given a cosmology and mass and energy content of the universe. One thus explores these interesting quantities. It is, however, known that the distribution of galaxies is biased away from the distribution of mass; exactly how much is a matter of some debate. The SDSS will allow these correlation functions to be measured and analyzed as a function of galaxy properties (e.g. magnitude, surface brightness, spectral type). If the redshift of the objects are known, either from spectroscopy or by photometric redshift techniques, one is able to compute the power spectrum directly. This often involves an SVD matrix calculation.

The essential procedure is to count the distance from each galaxy to every other galaxy, accumulating the statistics. This is an N^2 process (and higher order correlations equivalently higher order) though again metadata employing tree structures can cut the expense down to N log(N). The SVD matrix calculation is of order N^3. Neither program is particularly parallelizable, only made embarrassingly parallel by placing an arbitrary large angle cutoff. For the correlation function there is a further expensive step, that of the extensive randomization procedure that must be carried out in order to establish the statistical significance of the measurements.

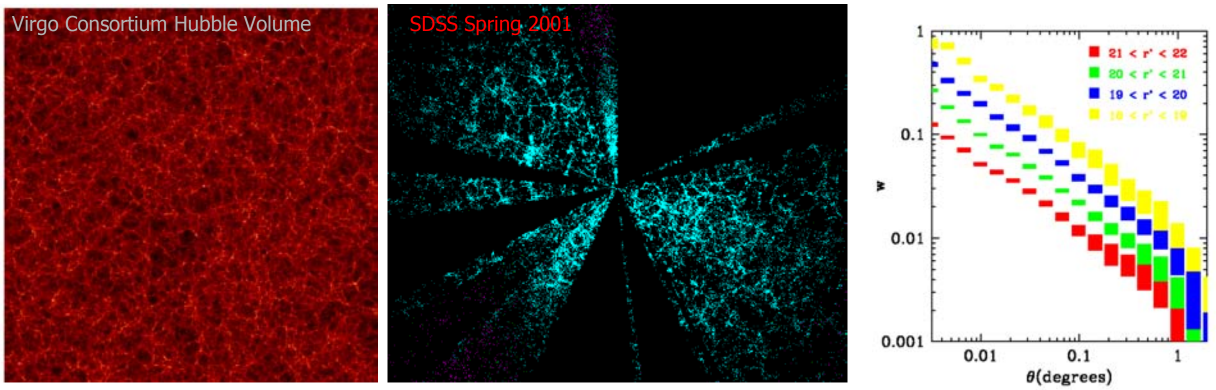

Figure 2: The correlation function. On the left panel, a numerical simulation of a Gigaparsec scale. In the middle, the SDSS redshift survey over a similar scale. On the right, the correlation function from the angular distribution of galaxies. The distribution of galaxies on the sky and in redshift contains cosmological information.

In both the correlation function measurement and the power spectrum measurement, the computation burden dominates the storage burden.

The correlation function challenge problem is an example of a catalog level analysis requiring large amounts of compute time. Each correlation function is computationally expensive, and very often the results are used as input to another layer of sophisticated codes. Correlation functions are an example of virtual data where a premium will be placed on locating existing materializations before requesting an expensive new materialization. If none can be found, then one enters the realm of resource discovery and job management.

3.3 The SDSS Challenge Problem 3: Weak Lensing

The analysis of weak lensing signals provides one of the very few direct measurements of mass available to astronomy. Weak lensing is caused by mass distorting the path of light from background objects; the objects tend to align in concentric circles. Tend is the operative phrase; weak lensing is an inherently statistical program. One use for weak lensing is to measure the masses of the clusters found by other means, and although the SDSS North is not deep enough for this, the SDSS South will be and in the meantime one can average together many clusters to achieve a signal. Another program, this one possible in the North, is to look at the random fluctuations in cells and use that to compute the power spectrum. The interesting thing here is that that is the power spectrum of the mass directly, without a worry about bias.

In order to do weak lensing, one must make suitably weighted second moments analysis of each image. Most of the algorithmic magic lies in the exact weighting, some in the details of the moment analysis.

This is an N^2 process in the number of pixels, which of course is much larger than the number of galaxies. In the SDSS data the objects detected have had atlas images cut out around them, which is very convenient for the weak lensing analysis, as one then must just make a pass through all of the galaxy atlas images.

Despite the heavy computational burden, this problem is weighted towards storage. The vast bulk of the atlas images that must be shepherded about dominates consideration of the problem. It is here that the bandwidth limitation raises its head, and the two distinct models of compute power local to the data and wide area grid computing must be balanced by a careful analysis.

The weak lensing challenge problem is a pixel level analysis that requires the moving of large amounts of data. Work will need to be done on data discovery, data and resource management.

4 The SDSS Experience: Towards Grid Enabled Astronomical Surveys

4.1 SDSS Data Replication

Several sites have expressed desire to mirror the SDSS data sets. Our database is SX, a custom layer on Objectivity. GDMP is nearly ideal for the mirroring. It implements a subscription model that is well matched to the problem at hand. It uses an efficient FTP, important when faced with 100 Gig scale transfers. It has transfer restart capability, very important when faced with the small pipe to Japan. The problem here will be to convince our colleagues to spend the energy to bring up Globus and GDMP. Unless it is very easy to install and run, astronomers will turn to existing, if non-optimal, tools such as shipping tapes.

4.2 The SDSS Pipeline Experience

4.2.1 Mapping the SDSS Pipelines into a GriPhyN Framework

The SDSS data processing is file oriented. While we use the object-oriented database SX, the objects are used only internally. The natural unit of data for most SDSS astronomers is the "field", one image of the sky, whose corresponding catalog has roughly 500 objects in it.

Data processing in SDSS is procedure oriented: start with raw data, run multi-stage processing programs, and save output files.

4.2.2 Description of the Abstracted SDSS Pipeline

The factory itself is the DP scripts that join together the pipelines. There are 3 generic stages of the factory at each and every pipeline:

INPUTS:

1. A plan file - defining which set of data are to be processed.
2. Parameter files - tuning parameters applicable to this particular set of data.
3. Input files - that are the products of upstream pipelines.
4. Environment variables - defining which versions of pipelines are being run.

```
1. Prep
        generate plan file containing, for example, root directories for input and output
        make relevant directories
        make sure the relevant data is available
        make relevant sym links
        locate space
        register the resources reserved in a flat file database
        Call submit

2. Submit
        generate a shell script the fires off the batch system submit
        Call ender

3. Ender
        periodically check status of a pipeline submit by looking for
                the files that should be created. The existence of the
                files is necessary and almost sufficient for the next step
                to succeed.
        after completion of the pipeline, run a quality control script,
                where various quantities (e.g., the number of galaxies/image)
                are given a sanity check.  if QC checks, then
        Call Prep for the next pipeline.
```

A complete "job" consists of the following steps:

1. Preparation. Stage any inputs that are needed. Create plan files, usually based on some rules about where directories are located.Make sure disk space is available. Inputs needed from previous pipelines are checked by looking for a "QC" file produced by any upstream pipelines.
2. Submit the pipeline job. The pipeline usually receives a few key inputs, such as the location of plan files. Return 0 or 1.
3. Run a status verify job that checks if the submitted job did complete.
4. Run a pipeline verify job to generate QC information and starts the follow-on pipeline. Return 0 or 1.
5. Run a "scrub" job that removes most files once they have been archived. The archiving is itself a "pipeline”.

OUTPUTS:

1. Output files - the data products themselves.
2. Log and error files - log and error files
3. Quality control files - identify outputs so fatally flawed that subsequent pipelines cannot run.
4. A single status flag – 0, proceed to next pipeline, 1, hand intervention required.

These are daisy chained: the first invocation of Prep takes as an argument how many of the following pipelines to run in series.

4.3 Southern Coadd Testbed

The analysis of weak lensing signals provides one of the very few direct measurements of mass available to astronomy, and there is great interest in pursing the analysis to the faintest magnitudes and widest sky coverage available. While the main SDSS The survey does not reach limiting fluxes faint enough to pursue true weak lensing maps, the SDSS Southern Survey does. In this survey, the same area of sky is imaged 20-60 times and the images added together to produce a image that reaches sqrt(N) times deeper in flux. The sqrt(N) is of interest: while the full complement of imaging data provides the maximal signal to noise, much of the improvement comes in the first handful of images and thus intermediate coadds are of interest.

We take the challenge of building weak lensing maps as the prototype Virtual Data problem in the Sloan Digital Sky Survey.

Data intensive science. The data sizes involved in the processing are hundreds of Gigs.

We take as the virtual data driver the science of analyzing the weak lensing map of the Southern coadded data. First things first: the southern coadd must be created, and that is only possible using existing data.

1) The Southern Survey Coadd
    - run full pipelines on incoming data
    - given an area on which to coadd,
        - find relevant reduced data
        - find disk space and compute power
        - extract relevant reduced data from long term storage
        - build mapping function from calibration data
        - perform coadd to create new reduced data
    - keep track of intermediate coadd catalogs and data

2) Run weak lensing analysis on the intermediate coadded south
    - locate atlas images on disk
    - compute optimal shape parameter
    - produce shape catalog

Clearly the right approach to the existing SDSS factory is to place it as a single coherent entity into the GriPhyN framework, as opposed to trying to put the individual pipelines into the framework.

Solving this problem provides avenues for future progress. One vision of the SDSS/NVO team is to spin the SDSS data on a compute cluster and allow for demand driven re-reduction with the everimproving versions of Photo. Of course, one cannot just run Photo, as it is the center of a long processing chain, but it is exactly this processing chain that must be made demand driven to solve the weak lensing map problem.

4.3.1 Data Sizes

The components of any given 200 sq-degree coadd of the south come to:

catalogs	measured parameters of all objects	10 Gig
atlas images	cutouts around all objects	15 Gig
binned sky	sky leftover after cutouts, binned 4x4	7 Gig
masks	regions of the sky not analyzed	7 Gig
calibration	a variety of calibration information	3 Gig
frames	corrected images	Nx200 Gig

4.3.2 Design Detail

```
1. given an area on which to coadd,
    a. Take user input
2. find relevant reduced data
    a.  Query either a db or a flat file for the the runs that have
        been observed and extract the list of runs that are of
        the right strip.
    b. Using that list, determine which runs overlap the given area.
        Involves the calculation of a spatial intersection.
    c. Apply cuts against data that are not of high enough quality to use
        (even though other parts of the run may be.)
    d. Estimate how much data is involved.
    Metadata: a list of runs and portions of runs involved

3. find disk space and compute power for input data, processing, and outputs
    a. Query the local network for available machines
    b. Query the local network for available storage
    c. Reserve the resources for a period of time
    Metadata: a list of machines

4. extract relevant reduced data from storage
    a. From some knowledge base, determine for each run if the data is:
        1) on archival disk
        2) on production disk
        3) on fast tape
        4) on slow tape
    b. Arrange to get the data off the media and onto the production
        machines
    Metadata: a list of  portions of runs and where they live on
        production disk
```

5. build mapping function from calibration data
    - a. An image is a 1361x2048 array of pixels. In general two images of the "same" piece of sky will be:
        - 1) slightly offset in both x and y
        - 2) slightly rotated
        - 3) have different small distortions as a function of x,y superimposed
        - 4) have different conversion between pixel value and energy flux
    - b. These translations, rotations, distortions, and scalings are calculateable from a set of calibration information that may or may not be kept as a header to the data itself.
    - c. Find the calibrations, and build the mapping function for each pixel
    - Metadata: computed mapping functions to be applied to the data

6. perform coadd to create new reduced data
    - a. Load a set of co-located images into memory
    - b. Apply mapping function
    - c. Perform median or average on stack of pixels (in truth: apply very clever algorithm to make maximal use of information and minimize noise)
    - d. Save output to disk.
    - Metadata: Location of output data

7. run full SDSS pipeline on new data set
    - a. arrange for all necessary input files to be in place
    - b. arrange for all 10 pipelines to run
    - c. watch success or failure of the pipelines
    - d. save resulting outputs to disk
    - Metadata: Location of output data

8. keep track of intermediate coadd catalogs and data
    - a. The output of 6. is to be preserved, as 7. can be done multiple times
    - b. The output of 7. is to be preserved
    - c. Each time new data comes in, the above is done.

9) Run weak lensing analysis on the intermediate coadded south
    - a) locate atlas images on disk
    - b) compute optimal shape parameter
    - c) produce shape catalog

5 Conclusion

The Southern Survey Coadd presents a local, manageable version of the full Virtual Data problem, and one which exhibits most of the major challenges of the full problem. Our collaboration will be ramping up to perform the Southern Survey over the next two years, and one way or another we will surmount the challenges and produce the weak lensing map. However, our activities during the process present the GriPhyN collaboration with an opportunity for alpha testing of tools, techniques, and ideas.